Published in final edited form as:
Genet Epidemiol. 2023 March ; 47(2): 152–166. doi:10.1002/gepi.22509.

Improved two-step testing of genome-wide gene-environment interactions

Eric S. Kawaguchi[1], **Andre E. Kim**[1], **Juan Pablo Lewinger**[1], **W. James Gauderman**[1]
[1]Department of Population and Public Health Sciences, University of Southern California, California, USA

Abstract

Two-step tests for gene-environment ($G \times E$) interactions exploit marginal SNP effects to improve the power of a genome-wide interaction scan (GWIS). They combine a screening step based on marginal effects used to 'bin' SNPs for weighted hypothesis testing in the second step to deliver greater power over single-step tests while preserving the genome-wide type I error. However, the presence of many SNPs with detectable marginal effects on the trait of interest can reduce power by 'displacing' true interactions with weaker marginal effects and by adding to the number of tests that need to be corrected for multiple testing. We introduce a new significance-based allocation into bins for step 2 $G \times E$ testing that overcomes the displacement issue and propose a computationally efficient approach to account for multiple testing within bins. Simulation results demonstrate that these simple improvements can provide substantially greater power than current methods under several scenarios. An application to a multi-study collaboration for understanding colorectal cancer (CRC) reveals a G×Sex interaction located near the SMAD7 gene.

Keywords

Cancer; Genome-wide; Linkage Disequilibrium; Power; Single nucleotide polymorphisms; Type I error

1 | INTRODUCTION

Identifying gene-environment interactions ($G \times E$) is critical for understanding how health is affected by both an individual's genetic background (G) and exposure to environmental factors (E). Genome-wide interaction scans (GWIS) test $G \times E$ interactions one-SNP-at-a-time by modeling the genotype, an environmental exposure, and the corresponding interaction term. Testing of the $G \times E$ is based on the significance of the interaction term, typically from a logistic (for a binary/disease trait), linear (for a quantitative trait), or Cox (1972) (for a survival trait) regression model. A standard one-step GWIS proceeds by testing each $G \times E$ interaction at significance level of $\alpha^* = 5 \times 10^{-8}$, common in genome-wide association studies (GWAS) or GWIS studies (Dudbridge & Gusnanto 2008). However, the

[*]**Correspondence:** Eric S. Kawaguchi, 1845 N. Soto St., Los Angeles, CA 90032, USA. ekawaguc@usc.edu.
Present Address
1845 N. Soto St., Los Angeles, CA 90032, USA.

statistical power to detect an interaction in a one-step GWIS is generally much lower than the power for detecting a genetic marginal effect in a GWAS.

Two-step tests for GWIS have been proposed to improve the power of a $G \times E$ analysis while controlling the FWER for disease (Gauderman, Zhang, Morrison, & Lewinger 2013; Hsu et al. 2012; Kooperberg & LeBlanc 2008; Murcray, Lewinger, Conti, Thomas, & Gauderman 2011; Murcray, Lewinger, & Gauderman 2009; Wang, Patel, Wason, & Newcombe 2021), quantitative (Paré, Cook, Ridker, & Chasman 2010; Zhang, Lewinger, Conti, Morrison, & Gauderman 2016), and time-to-event traits (Kawaguchi, Li, Lewinger, & Gauderman 2022). In all of these two-step procedures, independent information on $G \times E$ not captured by the standard $G \times E$ test is used to perform an initial screening (Step 1) to prioritize SNPs that are more likely to be involved in an interaction. These SNPs are formally tested for an interaction (Step 2) under a modified significance threshold α^*, thus reducing the multiple testing burden (Kooperberg & LeBlanc 2008; Murcray et al. 2009).

The marginal outcome-gene association statistic derived from modeling the outcome on each gene individually is a commonly-used screening statistic for quantitative (Zhang et al. 2016), binary/disease (Kooperberg & LeBlanc 2008), and time-to-event (Kawaguchi et al. 2022) traits. For case-control studies the exposure-gene association statistic, modeling the relationship between each gene on the exposure, can also be informative (Murcray et al. 2009). Methods that utilize both outcome-gene and exposure-gene associations in a case-control study have also been developed (Gauderman et al. 2013; Hsu et al. 2012; Murcray et al. 2011). A key requirement for validity of any two-step procedure is that the statistics used in Step 1 and Step 2 are asymptotically independent (Dai, Kooperberg, Leblanc, & Prentice 2012; Kawaguchi et al. 2022).

There are two widely-used procedures for prioritizing SNPs in Step-2 $G \times E$ testing after the Step-1 screening: subset (Kooperberg & LeBlanc 2008; Murcray et al. 2009) and weighted hypothesis testing (Ionita-Laza, McQueen, Laird, & Lange 2007). In subset testing, of the M total SNPs that are being scanned, only the $m << M$ SNPs that pass a significance threshold based on the screening statistic are tested in Step 2 using a standard $G \times E$ test. The significance threshold in step 2 for $G \times E$ discovery is calculated using a Bonferroni correction that is based on the number of SNPs that pass the screening $(\alpha^* = \alpha/m)$, which is much less stringent than the threshold used in a single step approach. A trade - off for a relaxed threshold is that SNPs that do not pass the step 1 screening will not be tested. An alternative approach that does not rely on a pass/no pass hard rule is weighted hypothesis testing. Here, SNPs are allocated into bins based on the magnitude of the screening statistic. Each bin has a corresponding bin-wise error rate (BWER) such that the sum across all bins does not exceed α. Top (higher priority) bins are allocated a larger fraction of α (see Section 2.1 for more detail), so that SNPs in those bins are tested at a more liberal significance threshold. Conversely, SNPs that are placed in lower-priority bins are tested at a much more stringent BWER. Unlike subset testing, every SNP is tested in Step 2 of the weighed approach; yet SNPs that are more likely to have an interaction based on the screening statistic will have a higher chance of being discovered. Although weighted hypothesis testing is often more powerful than subset testing (Gauderman et al. 2013; Ionita-Laza et al. 2007), there is no universally most powerful approach.

The motivation behind two-step hypothesis testing is that in the presence of a true $G \times E$ interaction effect, one can typically expect there to be marginal G effect on the outcome, which makes the marginal outcome-gene statistic useful for screening/ranking SNPs in Step 1 (e.g being placed in bin 1, the bin with the largest BWER). However, this only proves useful if not too many SNPs have sizeable marginal effects but no $G \times E$ interaction. This is often not the case in a GWIS where known “hits” from a prior GWAS provides a set of SNPs for which there is a strong marginal outcome-gene effect. For example, more than 140 GWAS-significant (marginal-effect) loci have been previously identified for colorectal cancer (Huyghe et al. 2019) and the majority of these likely do not exhibit a $G \times E$ interaction. This will result in a phenomena we refer to as “bin overcrowding”, where these SNPs will overcrowd the top bins in the weighted testing approach due to having non-zero marginal effects. Thus, even if a true $G \times E$ effect induces a non-zero marginal effect, it is competing against other known (or previously unknown) non-zero marginal effects which can force the true $G \times E$ effect to not be optimally tested (e.g. being placed/tested in a later bin with a stricter BWER). To avoid this loss of power, one can filter out these loci in advance. However, this approach is not ideal since it requires prior knowledge of loci with marginal effects and, more importantly, removes these SNPs from consideration for $G \times E$ testing.

The contribution of this paper is the improvement of two-step GWIS testing in two ways. First, we propose a simple yet effective approach to prioritize tests in the screening step that minimizes the potential power loss due to the presence of SNPs with a marginal effect but no interaction effect. Second, we show how additional accounting for correlation among SNPs in linkage disequilibrium (LD) in the testing step can further increase power. We demonstrate via simulation that this new two-step testing method yields greater power than its predecessors. We apply the approach to identify $G \times E$ interactions for colorectal cancer.

We describe current methods for testing $G \times E$ interactions using the two-step hypothesis testing framework and our proposed approach in Section 2. Simulation studies that compare the performance of two-step hypothesis testing procedures are given in Section 3 and an application to a GWIS for colorectal cancer is provided in Section 4. Lastly, in Section 5, we provide concluding remarks, limitations, and areas of future research.

2 | METHODS

Consider a gene-environment interaction study with a (continuous, binary, or time-to-event) trait/outcome Y, an environmental exposure of interest E, and M SNPs ($G_j, j = 1, \dots, M$) measured or imputed for each of the N subjects. To perform a GWIS, we assume M tests of $G \times E$ interaction with test statistics $\{T_j\}_{j=1}^{M}$ and corresponding p-values $\{p_j\}_{j=1}^{M}$ are computed. If for example the trait Y is quantitative, a standard one-step GWIS models the $G \times E$ interaction one-at-a-time by using the following:

$$E(Y \mid G_j, E) = \beta_{0j} + \beta_{G_j} G_j + \beta_E E + \beta_{G_j \times E}(G_j \times E) \tag{1}$$

for each of the M SNPs. For ease of exposition, we did not include subject-level covariates in the model, but in practice adjustment covariates like sex, age, and principal components capturing genetic ancestry should be considered . Each T_j corresponds to the test statistic for testing the null hypothesis $H_0: \beta_{G_j \times E} = 0$. An adjustment for multiple comparisons is applied to preserve the family-wise Type I error rate (FWER) at a prespecifi*j*ed significance level α (e.g., $\alpha^* = 5 \times 10^{-8}$ or $\alpha^* = \alpha / M$). However, this correction will lead to low power in detecting a $G \times E$ interaction.

2.1 | Two-step tests

Two-step methods have been developed to improve the power of GWIS(Kawaguchi et al. 2022; Kooperberg & LeBlanc 2008; Zhang et al. 2016). For a quantitative trait, the marginal outcome-gene association can be modeled using

$$E\left(Y \mid G_j\right) = \mu_{0j} + \mu_{G_j} G_j, \tag{2}$$

and S_j is the test statistic corresponding to $H_0: \mu_{G_j} = 0$. Model 2 has the form typically used to identify SNPs associated with the outcome in a standard GWAS. As with the one*j*-step GWIS, covariates should be carefully considered for inclusion in this model.

Two approaches to two-step testing are being widely used: subset testing (Kooperberg & LeBlanc 2008) and weighted hypothesis testing (Ionita-Laza et al. 2007). Subset testing: In subset testing, each of the M screening statistics is compared to a prespecified significance threshold α_0. Let $\mathscr{A}$ be the collection (subset) of indices of SNPs that pass the Step 1 screen (i.e. SNPs for which S_j is statistically significant at the α_0 level) with $|\mathscr{A}| = m$, where $|\mathscr{A}|$ represents the cardinality of set A. Then, for each $k \in \mathscr{A}$, the test statistic p-value p_k is compared against the Bonferroni-corrected significance level $\alpha^* = \alpha / m$, which is a less stringent significance level than in the standard single-step GWIS. Note that tests where $k \notin \mathscr{A}$ will never be tested – or equivalently tested against a significance level of $\alpha^* = 0$. This ensures that the overall Type I error rate is retained at α. Weighted testing: Instead of only testing a subset of hypotheses based on some dichotomous screening procedure, one can test all M test statistics in Step 2 using a weighted hypothesis test (Ionita-Laza et al. 2007) where the significance level assigned to each Step 2 SNP is based on the ordered (largest to smallest) absolute values of $\left\{S_j\right\}_{j=1}^{M}$ from Step 1. The rationale behind the weighted screening approach is that more promising SNPs - as measured by the screening statistic – are being tested at a less stringent Step-2 significance level . SNPs are assigned into B bins according to their M screening statistics. To control the FWER at level α, we derive bin-wise error rates (BWER) such that $\alpha_1 + \alpha_2 + \ldots + \alpha_B \leq \alpha$. A common choice is $\alpha_b = \alpha / 2^b$; then, for sufficiently large B,

$$\alpha_1 + \alpha_2 + \ldots + \alpha_B = \sum_{b=1}^{B} \frac{\alpha}{2^b} \approx \alpha\,.$$

By partitioning α in such a way, the overall Type I error rate is controlled at α while allocating a greater fraction of α to the op bins (i.e. those with the most promising SNPs). A Bonferroni-like correction can be used to preserve BWER by dividing α_b by $|\mathscr{A}_b| = m_b$

(size of the bin), where $\mathscr{A}_b$ is the set of indices for tests in bin b. We are now left with the non-trivial task of deciding how many SNPs should be allocated to each bin. Using a predetermined initial bin size, B_0, Ionita-Laza et al. (2007) suggested binning tests such that $m_b = 2^{b-1}B_0$. For example, the SNPs with the B_0 largest values of $|S_j|$ are placed in $\mathscr{A}_1$ and tested against a significance threshold of α_1/B_0, the next $2B_0$ SNPs are tested in bin 2 ($\mathscr{A}_2$) at level $\alpha_2/(2B_0)$ and so forth. Ionita-Laza et al. (2007) suggested setting $B_0 = 5$ so that $m_1 = 5$, $m_2 = 10$, $m_3 = 20$, $m_4 = 40$ … We refer to this Ionita-Laza et al. (2007) approach as rank-based (RB) binning to contrast with significance-based binning that we propose below. Two-step RB-weighted tests are generally more powerful than subset testing (Gauderman et al. 2013).

2.2 | Proposal # 1: Significance-based (SB) allocation of SNPs to bins in Step 1

Note that subset testing can be seen as a particular case of bin-based testing with only two bins with $\alpha_1 = \alpha$, $\mathscr{A}_1 = \mathscr{A}$, $\alpha_2 = 0$ and $\mathscr{A}_2 = \mathscr{A}^c$ where $\mathscr{A}$ is the subset of indices of SNPs that have a Step 1 p-value $< \alpha_0$. Unlike RB-weighted testing, where the size of bin 1 is set a priori, here the size of the bin is determined by the number of p-values in Step 1 that fall within the interval $(0, \alpha_0)$. Rather than creating two bins based on one p-value (e.g. α_0), one can allocate tests into B bins using a series of significance-level cutoffs. More specifically, defining $\tau = (\tau_1, \tau_2, \ldots, \tau_{B+1})$ as the set of significance level cutoffs, the collection of tests in bin b is $\mathscr{A}_b = \left\{ j : p_{S_j} \in [\tau_b, \tau_{b+1}) \right\}$. We refer to this type of screening as significance-based (SB) weighted testing, since SNP prioritization is based on the significance levels of the p-values corresponding to each screening statistic $\left(p_{S_j}\right)$. Thus, SB-weighted testing can be viewed as a hybrid between both subset and RB-weighted testing. While τ can be defined arbitrarily, we propose to set $\tau = (0, B_0/M, 3B_0/M, 7B_0/M, \ldots, 1)$ which, in expectation, corresponds to binnings that are identical to RB-weighted testing. However, bin sizes are not capped since p-values are used to determine bin allocation.

Assuming that the null hypothesis of no marginal association holds for all M SNPs, each p_{S_j} is uniformly distributed in the interval [0, 1]. In expectation, B_0 of the marginal outcome-gene screening statistics should have a p-value less than B_0/M, $2B_0$ of them should be in the interval $(B_0/M, 3B_0/M)$, and so forth. However, this relationship only holds under the null hypothesis of no outcome-gene associations. In practice, this assumption may not hold and will lead to a more-than-expected number of p-values to lie in the top bins. This overcrowding of the top bins (e.g., $\mathscr{A}_1$,$\mathscr{A}_2$) will reduce the power of the RB-weighted testing approach since bin sizes are capped and SNPs with true $G \times E$ effects may be pushed to a bin in which the interaction will be tested using a stricter threshold. As we will show in our numerical studies, the performance of RB-weighted testing is dependent on both the strength and number of non-null outcome-gene associations, which typically is unknown a priori. While overcrowding is still an issue with SB-weighted testing, we will show that the downstream effect on the testing step is reduced compared to RB-weighted testing.

2.3 | Proposal #2: Accounting for correlation among tests in Step 2

While two-step hypothesis testing generally provides increased power over the standard one-step GWIS, a Bonferroni-type correction is still used to control the FWER in Step 2. This adjustment tends to be overly conservative in situations where correlation is present as it is the case in association studies due to linkage disequilibrium (LD). Accounting for genetic correlation can provide additional power gains.

Permutation tests are seen as the gold standard by permuting the data in a way that simulates the null hypothesis while simultaneously maintaining the original correlation structure. However, in large association studies, permutation-based approaches are computationally prohibitive. Furthermore, it is not obvious how to extend permutation tests in the $G \times E$ setting due to the hierarchical structure of the model. Several more efficient methods to account for correlation between tests have been proposed. For example, Conneely and Boehnke (2007) developed a method (p_{ACT}) that attains the accuracy of permutation or simulation-based tests in much less computation time. Their approach, however, requires valid estimation of the covariance matrix, which is computationally prohibitive if applied genome wide. Another alternative is to replace the denominator of the Bonferroni correction, the total number of tests within a bin, with an estimate of the effective number of independent tests. Cheverud, Rutledge, and Atchley (1983), among others, have shown that the overall correlation among variables in a set can be captured by the variance of the eigenvalues derived from their correlation matrix (i.e. higher correlation among variables will lead to higher eigenvalue variance). Applying this correction genome wide is infeasible as it requires, first, the calculation of an $M \times M$ correlation matrix and, second, an eigendecomposition (e.g. principal components analysis) to derive the eigenvalues. We instead propose to apply the correction to the set of SNPs within each bin.

Let $\mathbf{G}$ be the $N \times M$ matrix of SNP genotypes (or imputed dosages) for N subjects and define $\mathbf{G}_{\mathscr{A}_b}$ to be the sub-matrix of $\mathbf{G}$ that corresponds to the indices in $\mathscr{A}_b(b = 1, \ldots, B)$. Let $\left\{\lambda_k^b\right\}_{k = 1}^K$ be the set of eigenvalues obtained by a principal component analysis (PCA) from the pair-wise SNP correlation matrix (i.e.. LD. matrix) of $\mathbf{G}_{\mathscr{A}_b}$, arranged in decreasing order. Gao, Starmer, and Martin (2008) proposed a simple and fast correction (simple$M$) that well approximates permutation-based corrections in both simulated and real data based on these eigenvalues

$$M_b^* = \min_x \left\{ \frac{\sum_{i = 1}^x \lambda_i^b}{\sum_{k = 1}^K \lambda_j^b} > C \right\}, \tag{3}$$

where C determines the percentage cutoff of explained variation. Gao et al. (2008) suggest setting $C = 0.995$ so that M_b^* corresponds to the minimum number of components needed to explain at least 99.5% of the variation in the data. Note that by design $\mathscr{A}_b$ grows larger in expectation as b increases. Thus the calculation of M_b^* for the lower-priority bins may still be computationally demanding. In practice, as shown in Section 3, we restrict the computation of M_b^* for the first seven bins and set the significance threshold for the later bins to be 0 since the threshold to be declared statistically significant in those upper bins is extremely stringent.

3 | SIMULATION STUDIES

We compare the performance of our two-step significance-based (SB) weighted testing procedure to 1) the standard one-step GWIS, and 2) the two-step rank-based (RB) weighted testing procedure. For both two-step RB and SB-testing, we compare the standard Bonferroni correction to the LD-adjusted Bonferroni correction using the simple$M$ approach (Gao et al. 2008) to account for LD within bins. As recommended by Gao et al. (2008), we set the tuning parameter $C = 0.995$.

Let **G** be an $N \times M$ genotype matrix for N individuals and M SNPs. In all of our simulations, we assume $M = 25,000$. e partition the $M = 25,000$ SNPs into blocks of 50 SNPs such that $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_{500}]$, where $\mathbf{G}_j$ is the jth block $N \times 50$ SNPs. Each $\mathbf{G}_j$ is simulated based on sampled minor allele frequencies (MAFs) and LD-matrices from the 1000 Genomes Project. For clarity, we denote G_j as the jth SNP and $\mathbf{G}_j$ as the jth block. Quantitative traits are simulated according to the following linear model:

$$Y = \beta_{G_{25}} G_{25} + \beta_E E + \beta_{G_{25} \times E}\left(G_{25} \times E\right) + \sum_{j \in \mathcal{G}} \beta_{G_j} G_j + \epsilon,$$

where $\epsilon \sim \mathcal{N}\left(0, \sigma_\epsilon^2 I\right)$ for some $\sigma_\epsilon^2 > 0$, E is the exposure variable (assumed to be binary) with $\Pr(E = 1) = 0.3$ and $\mathcal{G}$ corresponds to the set of SNPs that are only marginally associated with the outcome but have no $G \times E$ effect (G-only loci). By construction, the 25th SNP within block 1 ($\mathbf{G}_1$) has a true $G \times E$ effect on the outcome (i.e. the $G \times E$ locus). We refer readers to the Appendix for more information on the construction of **G** and the simulation setup.

The value of the parameters $\beta_{G_{25}}$, β_E, and $\beta_{G_{25} \times E}$ were set to achieve a predetermined R^2 for each term: $R_{G_{25}}^2 = 0.01$, $R_E^2 = 0.005$ and $R_{G_{25} \times E}^2 = 0.01$. Given a minor allele frequency of 0.231 for the $G \times E$ locus and setting $Var(Y) = 1$, these R^2 values correspond to $\beta_{G_{25}} \approx 0.06$, $\beta_E \approx -0.01$ and $\beta_{G_{25} \times E} \approx 0.37$. Each G-only locus was placed in a different block (i.e. 25th SNP in $\mathbf{G}_2$ 25th SNP in $\mathbf{G}_3$, etc.) so that these SNPs, and SNPs in LD, are expected to be prioritized in Step 1 and thus potentially affect power in identifying the true $G \times E$ effect due to bin overcrowding. For simplicity, we set the expected $R^2\left(R_G^2\right)$ for each of the G-only effects to be the same and vary the number of G-only effects (i.e.$|\mathcal{G}| = n_G$).

The testing statistics, T_j, are based on the hypotheses $H_0\colon \beta_{G_j \times E} = 0$ for $j = 1, \ldots, M$, where $\beta_{G_j \times E}$ is estimated based on the following one-SNP-at-a-time model

$$Y = \beta_j G_j + \beta_E E + \beta_{G_j \times E}\left(G_j \times E\right) + \epsilon,$$

where G_j denotes the jth SNP ($j = 1, \ldots , M$). For both two-step procedures (RB and SB-weighted), the screening statistics used in Step 1 are based on the test statistics corresponding to the marginal outcome-gene association model:

$$Y = \mu_0 + \mu_{G_j} G_j + \tilde{\epsilon} \quad j = 1, \ldots M\,.$$

The initial bin size B_0 was set to 5 SNPs for the RB-weighted approach and to an expectation of 5 SNPs assuming no genetic effects for the SB-weighted approach. For the latter, this corresponds to $\tau = (0, 5/25000, 15/25000, 35/25000, \ldots , 1)$. Under this scheme, we expect both RB and SB-weighted hypothesis testing to have comparable performance for detecting $G \times E$ loci when G-only effects are weak or absent. The FWER was set to $\alpha = 0.05$. As aforementioned, calculation of the effective number of independent tests for both two-step approaches will be computationally demanding for bins with larger number of SNPs. To avoid this, we restrict computation of M^* to the first seven bins and set the significance threshold in the later bins to 0.

Since SNPs in $\mathbf{G}_1$, the block with the causal $G \times E$ locus, are correlated, power is calculated as the number of times we reject the null hypothesis for any of the SNPs in $\mathbf{G}_1$ at the corresponding significance level. FWER is defined as rejecting the null hypothesis $H_0{:}\, \beta_{G_j \times E} = 0$ for $j = 51, \ldots , 25000$ at the corresponding significance level. Furthermore, we recorded the ranking of the $G \times E$ locus in terms of its Step 1 statistic as well as the smallest ranking of any loci in $\mathbf{G}_1$ in terms of its Step 1 statistic. Results are averaged over 5,000 Monte Carlo replications.

In our first set of simulations, we set $N = 2,000$ and $R^2_G = R^2_{G_{25}} = 0.01$ so that the marginal effects of the $G \times E$ locus and G-only loci are comparable. Power to detect the true $G \times E$ effect $\beta_{G_j \times E}$ with the standard one-step GWIS is approximately 45% (Figure 1 Panel A). When no G-only effects are present ($n_G = 0$), both approaches (RB and SB) have comparable power substantially higher than the standard one-step GWIS. The comparability of RB and SB-weighted testing in this scenario is expected since the binning of tests for both approaches should be nearly identical with no additional G-only effects. This is further supported by Figure 2 Panel A, which shows the distribution of the bin placement of the $G \times E$ locus. However, power for RB-weighted testing is sensitive to the number of G-only effects, dropping from ≈ 85% when only $n_G = 10$ G-only effects are present to ≈ 64% when $n_G = 80$ are present (Figure 1). This decrease in power is expected since the G-only loci, and their LD-regions, are being ranked higher than the $G \times E$ loci, and hence pushing the $G \times E$ loci into bins with more stringent significance thresholds in the testing step (Figure 2 Panel B). The loss in power for SB-weighted testing is less dramatic, which suggests that it is more robust to bin overcrowding than RB-weighted testing. Furthermore, adjusting for LD-based correlations in Step 2 provides a consistent increase in power across all scenarios (Figure 1).

In a second set of simulations (Figure 1 Panel B) we doubled the sample size N to 4,000 and halve the expected R^2 for each factor (i.e.$R^2_{G_{25}} = 0.005$, $R^2_E = 0.0025$, $R^2_{G_{25} \times E} = 0.005$, and $R_G^2 = 0.005$). Our results are consistent to what we have seen above which suggests that the power benefit of SB-weighted testing, coupled with the simple M procedure, can be achieved under more realistic effect sizes to what is commonly detected in GWAS and GWIS studies. As shown in Figure S1, the overall FWER is preserved at $<5\%$ for the RB and SB approaches in all of the above simulation scenarios.

In addition to increasing the number of marginal effects, we also evaluate both weighted testing methods when we vary both the size and magnitude of the G-only SNPs (Table 1). In this set of experiments, we fixed the percent of explained variation explained by the G-only SNPs to 40% and vary n_G from 10, 20, 40 to 80 such that the expected R^2 for each G-only SNP is $R_G^2 = 0.04$, 0.02, 0.01 and 0.005, respectively. We keep $R_{G_{25}}^2 = 0.01$ for the causal $G \times E$ locus so that we capture scenarios where the 'competition' in prioritizing the causal locus ranges from weak to strong. When R_G^2 is large relative to $R_{G_{25}}^2$, the screening test for the $G \times E$ locus has little-to-no chance of being in the top bins and and thus the Step-2 $G \times E$ hypothesis will always be tested at a more stringent significance threshold. Power for the standard one-step GWIS is unaffected by the magnitude and number of the G-only SNPs. The difference in power is largest between the SB and RB-weighted testing methods when a few number of 'strong' G-only SNPs ($n_G = 10, R_G^2 = 0.04$) are present. Here the median rank of the Step 1 statistic for the $G \times E$ locus = 61. In this scenario, RB-weighted testing has around 70% power of detecting the $G \times E$ locus whereas SB-weighted testing has power closer to 80%. On the other hand, power between both RB and SB-weighted testing are comparable when more but weaker G-only SNPs are present (around 80% power when $n_G = 80$ and $R_G^2 = 0.005$). Thus, both magnitude of effect and the number of G-only loci can greatly affect power for RB-weighted hypothesis testing. Conversely, SB-weighted screening is more robust to the size and magnitude of the G-only SNPs, with power hovering around $80 - 81\%$ in all four scenarios. This is also reflected in the bin placement of the simulated true $G \times E$ loci (Figure S2). The overall FWER is also preserved at $<5\%$ for the RB and SB approaches across these additional scenarios (Table S1).

It has been shown that the choice of B_0 affects power for weighted testing (Ionita-Laza et al. 2007; Lewinger et al. 2013). We investigate the performance of both RB and SB-weighted hypothesis testing under various initial bin sizes (Figure 3). For this simulation, we assume 80 G-only SNPs with $R_G^2 = 0.01$ are present ($R_{G_{25}}^2 = 0.01$, $R_E^2 = 0.005$ and $R_{G_{25} \times E}^2 = 0.01$) with $N = 2,000$ and $M = 25,000$. We see that RB-weighted screening is sensitive to B_0, with an increase in power as B_0 increases. The increase in power can be explained by the fact that we allow more tests to be included in each bin. Thus, even when comparably-sized G-only effects are competing with the causal $G \times E$ locus for bin placement, the probability of being in a smaller bin, and thus tested against a less stringent BWER, is higher. However, as others have explored, power can be sensitive to the choice of B_0, and larger values of B_0 can correspond to lower power when no G-only effects are present(Gauderman et al. 2013; Lewinger et al. 2013). In contrast, SB-weighted screening is relatively robust to the choice of B_0 used in creating the significance thresholds.

4 | APPLICATION TO COLORECTAL CANCER

We applied our proposed two-step approach to a genome-wide scan of gene-by-sex (G×Sex) interaction on the risk of colorectal cancer (CRC). The FIGI (Functionally Informed Gene-Environment Interaction) study is a multi-institutional collaborative effort to identify novel $G \times E$ interactions for CRC. Details of the FIGI study have been previously published (Huyghe et al. 2019; Schmit et al. 2019; Schumacher et al. 2015). In brief, epidemiological and genotype data were pooled from 61 cohort and case-control CRC studies from 3 large consortia - the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colorectal Transdisciplinary Study (CORECT), and Colon Cancer Family Registry (CCFR). Analyses include only individuals with complete exposure and covariate information, and were limited to individuals of European ancestry as determined by self-reported race and clustering of principal components with 1000 Genomes EUR sample. Details on genotyping, quality control, data collection and harmonization have been described previously (Hutter et al. 2012; Huyghe et al. 2019). Our analysis included $N = 89,304$ individuals (40,647 cases and 48,657 controls) and $M = 7,809,725$ imputed SNPs. Autosomal SNPs were imputed to the Haplotype Reference Consortium r1.1 reference panel via the Michigan Imputation Server (Das et al. 2016). Imputed SNPs were filtered based on a pooled MAF ≥ 1% and imputation accuracy of $R^2 > 0.80$.

Logistic regression models were used to model case/control status (Y) on each SNP allelic dosage (G), sex (E), and their interaction genomewide . The marginal outcome-gene associations, regressing Y on G genome-wide, were used as the screening statistics in Step 1. All models included the following set of adjustment covariates - age, study, and ancestry as defined by the first three principal components. Analyses using similar adjustment covariates were used in a recent investigation of G×Alcohol interaction in these data Jordahl et al. (2022).

Similar to the simulation study, bin significance threshold cutoffs for SB-weighted testing were based on cutoffs $\tau = (0, 5/M, 15/M, 35/M, \dots, 1)$. Bin flooding is apparent as bin 1 contains 3,795 SNPs (Figure 4) with a direct G-CRC association p-value $< 5/7,809,725$. Under the RB-weighted testing approach with $B_0 = 5$, these 3,759 SNPs fill the first 8 bins. For the SB approach, the estimated effective number of independent tests among the 3,759 SNPs is $M^* = 423$ based on the simple$M$ approach. Thus, each of the 3,795 $G \times E$ tests in bin 1 are tested against a significance threshold of $0.025/423$. One SNP (18:46458950:G:A) surpassed the threshold in bin 1. This SNP is located near the SMAD7 gene, which has been previously shown as a potential marker of colorectal cancer (Alidoust et al. 2022; Jiang et al. 2013; Stolfi et al. 2014; Thompson et al. 2009). This locus would not have been identified if the Bonferroni correction was used to correct for the number of tests in bin 1 (i.e. using significance threshold = 0.025 / 3,759). This result further supports the need to account for LD in the testing step. No statistically significant $G \times E$ effects were discovered using either the standard GWIS or the RB-weighted two-step approach. As a post-hoc analysis, we decided to run SB-weighted testing with $\tau = (0, 15/M, 45/M, \dots, 1)$. Under this specification of τ, bin 1 would contain SNPs that had a G-CRC association p-value $< 15/7,809,725$. From Figure 4 , we know that bin 1 would contain 3,795 (bin 1) + 895 (bin 2) = 4,690 SNPs. Figure S3 shows the results from the G-by-sex interaction scan, which still identifies SNP 18:46458950:G:A in bin 1 as a significant finding even under a more stringent significant threshold (0.025⁄490). Modifying the initial bin size to $B_0 = 15$ for the RB-weighted approach did not yield any statistically significant interactions.

5 | DISCUSSION

Two-step tests generally provide greater power than standard one-step approaches for genome-wide $G \times E$ discovery (Gauderman et al. 2017). We propose a novel approach, significance-based (SB) weighted hypothesis testing, that aims to address the shortcomings of its predecessors. In Step 1, we allocate SNPs into bins based on the significance of the screening statistics rather than on their rankings. Then, in step 2, to account for SNP-SNP correlations due to LD we control the bin-wise error rate based on an estimate of the effective number of independent tests. We show that SB-weighted testing is comparable to RB-weighted hypothesis testing when 1) no marginal G-only effects are present or 2) weak marginal G-only effect are present (i.e. little-to-no skew in the p-value distribution) and outperforms RB-weighted hypothesis testing when G-only loci flooding of the top bins exists. In addition, using an estimate of the effective number tests (simple$M$) to account for LD-based correlation among SNPs provides additional power for either of the two-step methods. We demonstrated our SB-weighted approach to identify a novel gene-sex interaction for colorectal cancer using data provided by the FIGI consortium.

In our simulation study, we also show that power for the RB-weighted hypothesis test is sensitive to both the number and magnitude of the step 1 screening statistics as well as the initial bin size. This corresponds to the number of marginal genetic effects that are associated with the outcome if the outcome-gene association statistic proposed by Kooperberg and LeBlanc (2008) is used in Step 1. These factors, among others, will negatively affect the prioritization of the causal ($G \times E$) loci in the testing step since the number of tests per bin are capped a priori. Alternatively, the use of significance cutoffs in the SB-weighted approach is robust to both the number and magnitude of the step 1 statistics and has greater power over RB-weighted testing when there is competition in bin prioritization. Adjusting for the effective number of tests using the simple$M$ method of Gao et al. (2008) yields additional improvements in power while preserving the FWER.

Our two-step hypothesis testing procedure identified a new locus near SMAD7 that interacts with biological sex to modulate CRC risk. Further studies should examine the potential mechanisms through which this newly discovered locus impacts CRC risk. We note that G×Sex interaction for this locus would not have been found using either the standard one-step GWIS or the RB-weighted testing approach.

We envision several directions to further explore SB-weighted testing. First, the marginal outcome-gene association was used as the screening statistic in Step 1. However, SB-weighted testing should be valid using any screening statistic in Step 1 as long as it is independent of the testing statistic in Step 2 (Dai et al. 2012). Further investigation into the performance of SB-weighted testing is warranted under different phenotypes, sampling designs, and screening statistics. The binning of tests was performed by developing significance-level based cutoffs of the p-values in Step 1. Our rationale for this approach was to, in expectation, develop bins that were similar to the rank-based binnings proposed by Ionita-Laza et al. (2007). However, one can bin tests arbitrarily based on the p-value distribution of the Step 1 screening statistics or on the distribution of the screening statistics themselves. Alternative approaches for both screening and test binning warrant future investigation.

LD was taken into account through estimating the effective number of independent tests. While we explored the performance of the simple$M$ procedure proposed by Gao et al. (2008), several other estimators have been suggested (Cheverud 2001; Galwey 2009; J. Li & Ji 2005; M.-X. Li, Yeung, Cherny, & Sham 2012; Nyholt 2004). Alternative approaches to adjust p-values within bins (or the testing subset) can be adopted (Conneely & Boehnke 2007). While genetic correlation within bins have been considered, between bin LD can be present as well. Methods to account for both within- and between-bin genetic correlation may further improve power and should be investigated further. Alternatively, one may instead control the false discovery rate (Benjamini & Hochberg 1995), the expected proportion of discoveries that are false, rather than the FWER. Future research is needed into developing FDR-based two-step hypothesis tests.

We have demonstrated that the current two-step $G \times E$ testing framework can be greatly improved by 1) incorporating a more robust binning procedure for the screening step (Step 1) and 2) taking SNP LD into account when setting the Step-2 significance thresholds. We have demonstrated, by simulation and application to a colorectal cancer study, that significance-based binning of Step-1 tests and LD-correction of thresholds in Step 2 have the potential to discover novel $G \times E$ interactions for a complex trait. Additional examinations of the binning approach and the method to account for LD to further improve 2-step GWIS power are warranted.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The work was partially supported by NIH grants T32ES013678, P01CA196569, R01CA201407, and P30ES007048. These awards had no influence over the experimental design, data analysis or interpretation, or writing the manuscript.

Abbreviations:

BWER	bin-wise error rate
FWER	family-wise error rate
GWAS	genome-wide association study
GWIS	genome-wide interaction scans
LD	linkage disquilibrium
RB	rank-based

SB	significance-based
SNP	single nucleotide polymorphism

References

Alidoust M, Hamzehzadeh L, Khorshid Shamshiri A, Afzaljavan F, Kerachian MA, Fanipakdel A, … others (2022). Association of smad7 genetic markers and haplotypes with colorectal cancer risk. BMC medical genomics, 15(1), 1–9. doi: 10.1186/s12920-021-01150-3 [PubMed: 34980126]

Benjamini Y, & Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Cheverud JM (2001). A simple correction for multiple comparisons in interval mapping genome scans. Heredity, 87(1), 52–58. doi: j.1365-2540.2001.00901.x [PubMed: 11678987]

Cheverud JM, Rutledge J, & Atchley WR (1983). Quantitative genetics of development: genetic correlations among age-specific trait values and the evolution of ontogeny. Evolution, 895–905. doi: 10.1111/j.1558-5646.1983.tb05619.x [PubMed: 28563537]

Conneely KN, & Boehnke M (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. The American Journal of Human Genetics, 81(6), 1158–1168. doi: 10.1086/522036 [PubMed: 17966093]

Cox DR (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x

Dai JY, Kooperberg C, Leblanc M, & Prentice RL (2012). Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. Biometrika, 99(4), 929–944. doi: 10.1093/biomet/ass044 [PubMed: 23843674]

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, ... others (2016). Next-generation genotype imputation service and methods. Nature genetics, 48(10), 1284–1287. doi: 10.1038/ng.3656 [PubMed: 27571263]

Dudbridge F, & Gusnanto A (2008). Estimation of significance thresholds for genomewide association scans. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 32(3), 227–234. doi: 10.1002/gepi.20297

Galwey NW (2009). A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 33(7), 559–568. doi: 10.1002/gepi.20408

Gao X, Starmer J, & Martin ER (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 32(4), 361–369. doi: 10.1002/gepi.20310

Gauderman WJ, Mukherjee B, Aschard H, Hsu L, Lewinger JP, Patel CJ, ... others (2017). Update on the state of the science for analytical methods for gene-environment interactions. American journal of epidemiology, 186(7), 762–770. doi: 10.1093/aje/kwx228 [PubMed: 28978192]

Gauderman WJ, Zhang P, Morrison JL, & Lewinger JP (2013). Finding novel genes by testing g× e interactions in a genome-wide association study. Genetic epidemiology, 37(6), 603–613. doi: 10.1002/gepi.21748 [PubMed: 23873611]

Hsu L, Jiao S, Dai JY, Hutter C, Peters U, & Kooperberg C (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. Genetic Epidemiology, 36(3), 183–194. doi: 10.1002/gepi.21610 [PubMed: 22714933]

Hutter CM, Chang-Claude J, Slattery ML, Pflugeisen BM, Lin Y, Duggan D, ... others (2012). Characterization of gene–environment interactions for colorectal cancer susceptibility loci. Cancer research, 72(8), 2036–2044. doi: 10.1158/0008-5472.CAN-11-4067 [PubMed: 22367214]

Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, ... others (2019). Discovery of common and rare genetic risk variants for colorectal cancer. Nature genetics, 51(1), 76–87. doi: 10.1038/s41588-018-0286-6 [PubMed: 30510241]

Ionita-Laza I, McQueen MB, Laird NM, & Lange C (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. The American Journal of Human Genetics, 81(3), 607–614. doi: 10.1086/519748 [PubMed: 17701906]

Jiang X, Castelao JE, Vandenberg D, Carracedo A, Redondo CM, Conti DV, ... others (2013). Genetic variations in smad7 are associated with colorectal cancer risk in the colon cancer family registry. PloS one, 8(4), e60464. doi: 10.1371/journal.pone.0060464 [PubMed: 23560096]

Jordahl KM, Shcherbina A, Kim AE, Su Y-R, Lin Y, Wang J, ... others (2022). Beyond gwas of colorectal cancer: Evidence of interaction with alcohol consumption and putative causal variant for the 10q24. 2 region. Cancer Epidemiology, Biomarkers & Prevention, 31(5), 1077–1089. doi: 10.1158/1055-9965.EPI-21-1003

Kawaguchi ES, Li G, Lewinger JP, & Gauderman WJ (2022). Two-step hypothesis testing to detect gene-environment interactions in a genome-wide scan with a survival endpoint. Statistics in Medicine, 41(9), 1644–1657. doi: 10.1002/sim.9319 [PubMed: 35075649]

Kooperberg C, & LeBlanc M (2008). Increasing the power of identifying gene× gene interactions in genome-wide association studies. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 32(3), 255–263. doi: 10.1002/gepi.20300

Lewinger JP, Morrison JL, Thomas DC, Murcray CE, Conti DV, Li D, & Gauderman WJ (2013). Efficient two-step testing of gene-gene interactions in genome-wide association studies. Genetic epidemiology, 37(5), 440–451. doi: 10.1002/gepi.21720 [PubMed: 23633124]

Li J, & Ji L (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity, 95(3), 221–227. doi: 10.1038/sj.hdy.6800717 [PubMed: 16077740]

Li M-X, Yeung JM, Cherny SS, & Sham PC (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Human genetics, 131(5), 747–756. doi: 10.1007/s00439-011-1118-2 [PubMed: 22143225]

Murcray CE, Lewinger JP, Conti DV, Thomas DC, & Gauderman WJ (2011). Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genetic epidemiology, 35(3), 201–210. doi: 10.1002/gepi.20569 [PubMed: 21308767]

Murcray CE, Lewinger JP, & Gauderman WJ (2009). Gene-environment interaction in genome-wide association studies. American journal of epidemiology, 169(2), 219–226. doi: 10.1093/aje/kwn353 [PubMed: 19022827]

Nyholt DR (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. The American Journal of Human Genetics, 74(4), 765–769. doi: 10.1086/383251 [PubMed: 14997420]

Paré G, Cook NR, Ridker PM, & Chasman DI (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. PLoS Genet, 6(6), e1000981. doi: 10.1371/journal.pgen.1000981 [PubMed: 20585554]

Schmit SL, Edlund CK, Schumacher FR, Gong J, Harrison TA, Huyghe JR, ... others (2019). Novel common genetic susceptibility loci for colorectal cancer. JNCI: Journal of the National Cancer Institute, 111(2), 146–157. doi: 10.1093/jnci/djy099 [PubMed: 29917119]

Schumacher FR, Schmit SL, Jiao S, Edlund CK, Wang H, Zhang B, ... others (2015). Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nature communications, 6(1), 1–7. doi: 10.1038/ncomms8138

Stolfi C, De Simone V, Colantoni A, Franze E, Ribichini E, Fantini M, ... others (2014). A functional role for smad7 in sustaining colon cancer cell growth and survival. Cell death & disease, 5(2), e1073–e1073. doi: 10.1038/cddis.2014.49 [PubMed: 24556688]

Thompson CL, Plummer SJ, Acheson LS, Tucker TC, Casey G, & Li L (2009). Association of common genetic variants in smad7 and risk of colon cancer. Carcinogenesis, 30(6), 982–986. doi: 10.1093/carcin/bgp086 [PubMed: 19357349]

Wang J, Patel A, Wason JM, & Newcombe PJ (2021). Two-stage penalized regression screening to detect biomarker-treatment interactions in randomized clinical trials. Biometrics. doi: 10.1111/biom.13424

Zhang P, Lewinger JP, Conti D, Morrison JL, & Gauderman WJ (2016). Detecting gene–environment interactions for a quantitative trait in a genome-wide association study. Genetic epidemiology, 40(5), 394–403. doi: 10.1002/gepi.21977 [PubMed: 27230133]

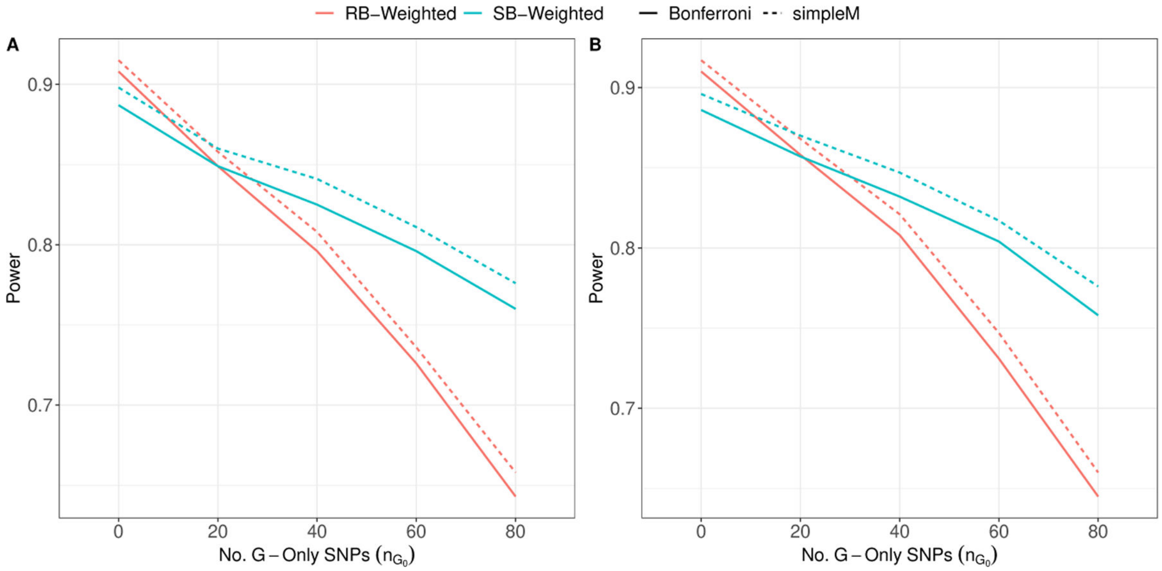

FIGURE 1.

Estimated power when n_G G-only effects are present ($n_G \in \{10, 20, 40, 80\}$) that each explain $R_G^2 \times 100\%$ of the variation. RB-Weighted: Rank-based weighted hypothesis testing using with initial bin size B_0 in Step 1; SB-Weighted: Significance-based weighted hypothesis testing using $\tau = (0, B_0/25000, 3B_0/25000, \ldots, 1)$ as the p-value cutoffs in Step 1. Bonferroni: Standard Bonferroni correction within bin; simpleM: The simpleM procedure proposed by Gao et al. (2008) with $C = 0.995$. Results are averaged over 5,000 simulations. Panel A:$R_{G25}^2 = R_{G25 \times E}^2 = R_G^2 = 0.01$, $R_E^2 = 0.005$, $N = 2,000$; Panel B:$R_{G25}^2 = R_{G25 \times E}^2 = R_G^2 = 0.005$, $R_E^2 = 0.0025$, $N = 4,000$.

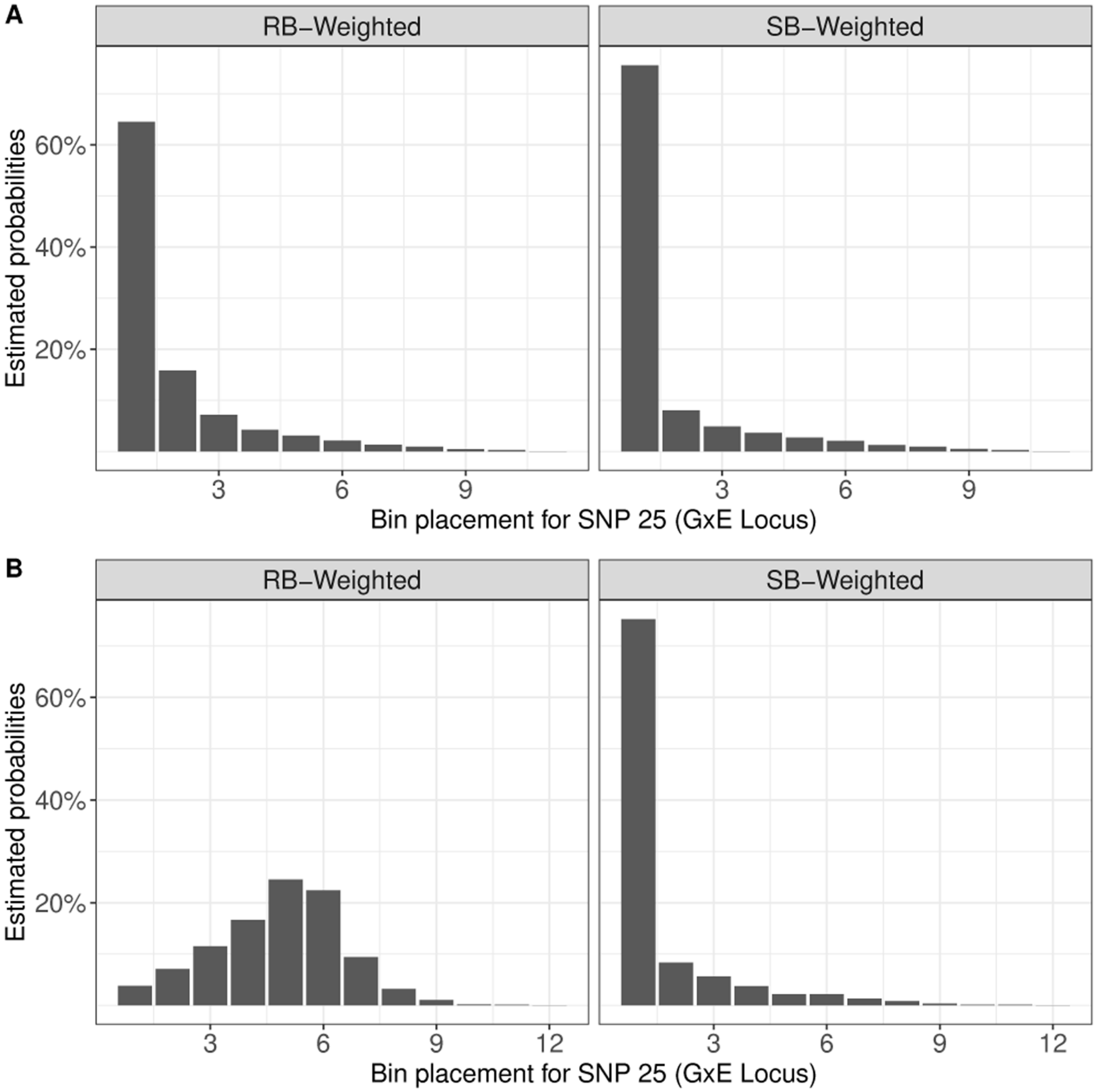

FIGURE 2.

Bar chart of bin placement for the 25th SNP (i.e. $G \times E$ locus) in Step 1 over 5,000 simulations. RB-Weighted: Rank-based weighted hypothesis testing using with initial bin size B_0 in Step 1; SB-Weighted: Significance-based weighted hypothesis testing using $\tau = (0, B_0/25000, 3B_0/25000, \ldots, 1)$ as the p-value cutoffs in Step 1. Simulation parameters: $R^2_{G25} = R^2_{G25 \times E}$, $R^2_E = 0.005$, $N = 2,000$, $M = 25,000$. Panel A) $n_G = 0$ G-only SNPs; Panel B: $n_G = 80$ G-only SNPs with $R^2_G = 0.01$.

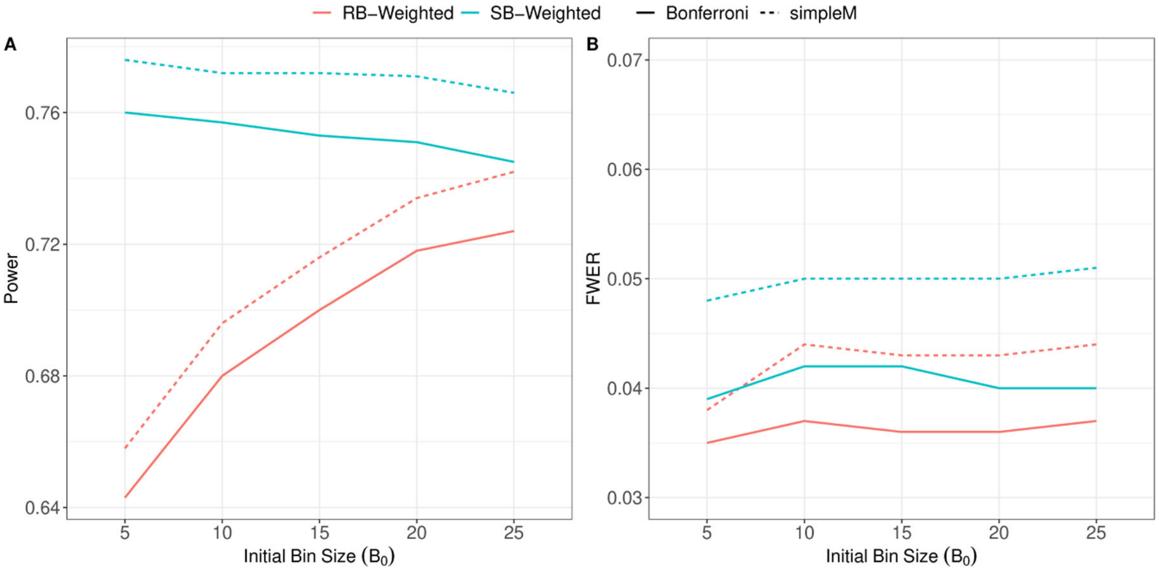

FIGURE 3.

Estimated power and FWER as a function of the initial bin size B_0. 80 G-only effects are present with $R_G^2 = 0.01$. RB-Weighted: Rank-based weighted hypothesis testing using with initial bin size B_0 in Step 1; SB-Weighted: Significance-based weighted hypothesis testing using $\tau = (0, B_0/25000, 3B_0/25000, \ldots, 1)$ as the p-value cutoffs in Step 1. Bonferroni: Standard Bonferroni correction within bin or subset; simpleM: The simpleM procedure proposed by Gao et al. (2008) with $C = 0.995$. Results are averaged over 5000 simulations.

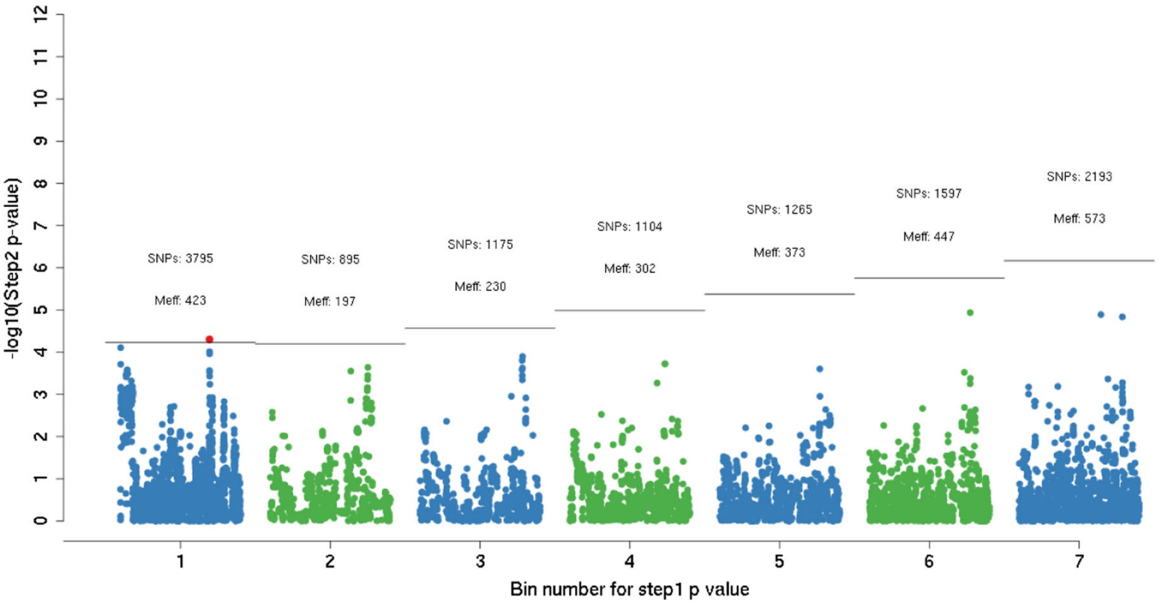

FIGURE 4.

Results from the G-by-sex interaction scan using the SB-weighted testing approach applied to the FIGI consortium data ($N = 89, 304$, $M = 7, 809, 725$). x-axis: Bins are based on the marginal outcome-gene association statistic (e.g. SNPs that have a Step 1 statistic $< 5/M$ are included in bin 1). y-axis: p-value of the $G \times E$ association provided by the GWIS (on the $-\log_{10}$ scale). Number of SNPs in each bin as well as the effective number of independent SNPs (Meff) using the simple$M$ approach are included. Horizontal line indicates the threshold the Step 2 p-value must cross to be statistically significant, maintaining the overall FWER=0.05. Only SNPs in the first 7 bins are shown in this figure.

TABLE 1

Estimated power when n_G G-only effects are present ($n_G \in \{10, 20, 40, 80\}$) and the total amount of variation explained is fixed at 40%$\left(R_G^2 = 0.4/n_G\right)$. One-step GWIS: The standard one-step GWIS. RB-Weighted: Rank-based weighted hypothesis testing proposed by Ionita-Laza et al. (2007) with $B_0 = 5$; SB-Weighted: Our proposed significance-based weighted hypothesis testing with $\tau = (0, 5/25000, 15/25000, \ldots, 1)$ as the p-value cutoffs. Bonferroni: Standard Bonferroni correction within bin; simple$M$: The simpleM procedure proposed by Gao et al. (2008) with $C = 0.995$. Results are averaged over 5000 simulations.

n_G =	**10**	**20**	**40**	**80**
$R_G^2 =$	0.04	0.02	0.01	0.005
One-step GWIS	0.439	0.437	0.443	0.442
RB-Weighted				
Bonferroni	0.694	0.670	0.726	0.801
simple$M$	0.713	0.688	0.736	0.812
SB-Weighted				
Bonferroni	0.814	0.797	0.796	0.812
simple$M$	0.829	0.811	0.812	0.826